Robust Estimation of Dimension Reduction Space

Cizek, P.; Härdle, W.K.

Publication date:
2005

No. 2005–31

ROBUST ESTIMATION OF DIMENSION REDUCTION SPACE

By Pavel Čížek, Wolfgang Härdle

February 2005

ISSN 0924-7815

Robust estimation of dimension reduction space

P. Čížek [a] and W. Härdle [b]

[a] *Department of Econometrics & OR, Tilburg University, P.O.Box 90153, 5000 LE, Tilburg, The Netherlands*

[b] *Institut für Statistik und Ökonometrie, Humboldt-Universität zu Berlin, Spandauer Str. 1, D-10178 Berlin, Germany*

Abstract

Most dimension reduction methods based on nonparametric smoothing are highly sensitive to outliers and to data coming from heavy-tailed distributions. We show that the recently proposed methods by Xia et al. (2002) can be made robust in such a way that preserves all advantages of the original approach. Their extension based on the local one-step M-estimators is sufficiently robust to outliers and data from heavy tailed distributions, it is relatively easy to implement, and surprisingly, it performs as well as the original methods when applied to normally distributed data.

JEL codes: C14, C20

Key words: Dimension reduction, Nonparametric regression, M-estimation

Email addresses: P.Cizek@uvt.nl (P. Čížek), haerdle@wiwi.hu-berlin.de (W. Härdle).

1 Introduction

In regression, we aim to estimate the regression function, which describes the relationship between a dependent variable $y \in \mathbb{R}$ and explanatory variables $X \in \mathbb{R}^p$. This relationship can be, without prior knowledge and with full generality, modelled nonparametrically, but an increasing number of explanatory variables makes nonparametric estimation suffer from the curse of dimensionality. There are two main approaches to deal with a large number of explanatory variables: to assume a simpler form of the regression function (e.g., its additivity) or to reduce the dimension of the space of explanatory variables. The latter, more general approach received a lot of attention recently; see Li (1991) and Xia et al. (2002), for instance. Our first aim is to study the latter approach and examine its sensitivity to heavy-tailed and contaminated data, which can adversely influence both parametric and nonparametric estimation methods (see Čížek, 2004, and Sakata and White, 1998, for evidence in financial data). Further, we propose robust and computationally feasible modifications of Xia et al. (2002)'s methods and study their behavior by means of Monte Carlo experiments.

A dimension-reduction (DR) regression model can be written as

$$y = g(B_0^\top X) + \varepsilon, \tag{1}$$

where g is an unknown smooth link function, B_0 represents a $p \times D$ orthogonal matrix, $D \leq p$, and $E(\varepsilon|X) = 0$ almost surely. Hence, to explain the dependent variable y, the space of p explanatory variables X can be reduced to a D-dimensional space given by B_0 (for $D = p$, the standard nonparametric regression model results). The vectors of B_0 are called directions in this con-

text. The dimension reduction methods aim to find the dimension D of the DR space and the matrix B_0 defining this space.

Recently, Xia et al. (2002) proposed the minimum average variance estimator (MAVE), which improves in several aspects over other existing estimators, such as sliced inverse regression (SIR) by Li (1991). First, MAVE does not need undersmoothing when estimating the link function g to achieve a faster rate of convergence. Second, MAVE can be applied to many models including time series data and readily extended to other related problems such as classification (Antoniadis et al., 2003) and functional data analysis (Amato et al., 2005). Finally, the MAVE approach renders generalizations of some other nonparametric methods; for example, Xia's outer product of gradients (OPG) estimator extends the average derivative estimator (ADE) of Härdle and Stoker (1989) to multi-index models.

Despite many features, MAVE does not seem to be robust to outliers in the dependent variable y since it is based on local least-squares estimation (for evidence, see Rousseeuw and Leroy, 2003, in parametric and Čížek, 2004, in nonparametric regression). Similar sensitivity to outliers in the space of explanatory variables X (so-called leverage points), was observed and remedied in the case of the sliced inverse regression (SIR) by Gather et al. (2001). At the same time, the robustness of DR methods is crucial since analyzed data are typically highly dimensional, and as such, are difficult to check and clean. Therefore, because of many advantages that MAVE possess, we address its low robustness to outlying observations and propose ways to improve it without affecting main strengths of MAVE. Additionally, we also employ and generalize OPG because, despite being inferior to MAVE, it provides an easy-to-implement and fast-to-compute method that could prove preferable in some

applications (especially if dimension D is expected to be very small).

The rest of the paper is organised as follows. In Section 2, we describe both the MAVE and OPG methods and discuss their sensitivity to outliers. Robust enhancements of these methods are proposed in Section 3. Finally, we compare all methods by means of simulations in Section 4.

2 Estimation of dimension reduction space

In this section, we first present MAVE and OPG (Sections 2.1 and 2.2) as well as a procedure for determining the effective DR dimension by means of cross validation (Section 2.3). At the end of the section, we will motivate our concerns about robustness of these methods (Section 2.4).

2.1 The MAVE method

Let d represent the working dimension, $1 \leq d \leq p$. For a given number d of directions B in model (1), Xia et al. (2002) proposed to estimate B by minimizing the unexplained variance $E\{y - E(y|X)\}^2 = E\{y - g(B^\top X)\}^2$, where the unknown function g is locally approximated by a linear function; that is, $g(B^\top X_0) \approx a_0 + b_0^\top B^\top (X - X_0)$ around some X_0. The novel feature of MAVE is that one minimizes simultaneously with respect to directions B and coefficients a_0 and b_0 of the local linear approximation. Hence, given a sample

$(X_i, y_i)_{i=1}^n$ from (X, y), MAVE estimates B by minimizing

$$\min_{\substack{B:B^\top B=I_p \\ a_j,b_j,j=1,\ldots,n}} \sum_{i=1}^{n}\sum_{j=1}^{n}[y_i - \{a_j + b_j^\top B^\top (X_i - X_j)\}]^2 w_{ij}, \tag{2}$$

where w_{ij} are weights describing the local character of linear approximation. Initially, weights at any point X_0 are given by a multidimensional kernel function K_h, where h refers to a bandwidth: $w_{i0} = K_h(X_i - X_0)\{\sum_{i=1}^n K_h(X_i - X_0)\}^{-1}; i = 1, \ldots, n$. Additionally, once we have an estimate $\hat{B}$ of the DR space, it is possible to iterate using weights based on distances in the reduced space: $w_{i0} = K_h\{\hat{B}^\top(X_i - X_0)\}[\sum_{i=1}^n K_h\{\hat{B}^\top(X_i - X_0)\}]^{-1}$. Iterating until convergence results in a refined MAVE, which is the estimator we understand in the rest of the paper under MAVE.

Xia et al. (2002) also proposed a non-trivial iterative estimation procedure based on repeating two simpler optimizations of (2): one with respect to a_j, b_j given an estimate $\hat{B}$ and another with respect to B given estimates $\hat{a}_j, \hat{b}_j$. This computational approach greatly simplifies and speeds up estimation.

2.2 The OPG method

Based on the MAVE approach, Xia et al. (2002) also generalised ADE of Härdle and Stoker (1989) to multi-index models. Instead of using a moment condition for the gradient of the regression function g in model (1), $E\{\nabla g(X)\} = 0$, the average outer product of gradients (OPG) is used: $\Sigma = E\{\nabla g(X)\nabla^\top g(X)\}$. It can be shown that the DR matrix B consists of the d

eigenvectors corresponding to the d largest eigenvalues of Σ. Thus, recalling that local linear regression solves at point X_j

$$\min_{a_j,b_j} \sum_{i=1}^{n} [y_i - \{a_j + b_j^\top (X_i - X_j)\}]^2 w_{ij}, \tag{3}$$

we can estimate Σ by $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \hat{b}_j^\top \hat{b}_j$, where $\hat{b}_j$ are estimates of b_j from (3). Hence, OPG consists in estimating $\hat{\Sigma}$ and determining its d eigenvectors with largest eigenvalues. The choice of weights w_{ij} can be done in the same way as for MAVE.

Although OPG does not exhibit the convergence rate of MAVE, it is easy to implement, fast to compute, and can be flexibly combined with robust estimation methods as shown in Section 3. Moreover, our simulations show that it can perform as well as MAVE if just one or two directions, $d \leq 2$, are of interest.

2.3 Dimension of effective reduction space

The described methods can estimate the DR space for a pre-specified dimension d. To determine d, Xia et al. (2002) extend the cross-validation (CV) approach of Yao and Tong (1994) and estimate d by $\hat{d} = \text{argmin}_{0 \leq d \leq p} CV(d)$, where

$$CV(d) = \sum_{j=1}^{n} \left[y_j - \sum_{i=1, i \neq j}^{n} \frac{y_i K_h\{\hat{B}^\top (X_i - X_j)\}}{\sum_{i=1, i \neq j}^{n} K_h\{\hat{B}^\top (X_i - X_j)\}} \right]^2 \tag{4}$$

for $d > 0$ and $CV(0) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2$ to incorporate the possibility of y and X being independent. Note that this CV criterion can be also used to select bandwidth h, which results in two-dimensional CV over d and h. Although we

use this time-demanding strategy in our simulations, in practice it is possible to consistently select bandwidth h^* in the DR space for the largest dimension $d = p$ and employ this bandwidth for all other dimensions.

2.4 Robustness of dimension reduction

The complexity of highly dimensional data processed by means of dimension reduction methods requires estimation methodology robust to data contamination, which can arise from miscoding or heterogeneity not captured or presumed in a model. In nonparametric regression, even data coming from a heavy-tailed distribution can exhibit effects similar to data contamination. Since both MAVE, OPG, and CV are based on least-squares criteria, their sensitivity to outlying observations can be rather high. Here we discuss possible effects of a single outlying observation on the estimation results, that is on $\hat{a}_j, \hat{b}_j$, and $\hat{B}$, and on the CV selection of bandwidth h and dimension d. At the end of the section, we demonstrate the effects of an outlier on a simple example.

Considering OPG and local parametric regression (3), the estimates $\hat{a}_j$ and $\hat{b}_j$ are just a linear combination of values y_i. Since the weights w_{ij} are independent of y_i for a given bandwidth h, even a single outlying value $y_i, |y_i| \to \infty$, can arbitrarily change the estimated coefficients $\hat{a}_j$ and $\hat{b}_j$ around X_j if h is sufficiently large. This effect then influences matrix $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \hat{b}_j^\top \hat{b}_j$. In the case of MAVE defined by (2), the situation is more complicated since the local linear approximation of the link function given by a_j and b_j can adjust simultaneously with directions B. In general, it is not possible to explicitly state, which parameters will be affected and how, but it is likely that the effect

Table 1

Optimal CV bandwidth for dimension d and a data set with one additive outlier c.

Simulated data with an additive outlier						
Dimension d	1	2	4	6	8	10
$c = 0$	0.61	1.28	1.65	1.28	1.28	1.65
$c = 50$	0.37	1.00	1.65	1.28	1.65	2.12
$c = 250$	0.47	1.28	2.12	2.12	0.47	0.61
$c = 500$	0.47	0.78	0.22	0.29	0.37	0.61
$c = 1000$	0.47	0.61	0.22	0.37	0.47	0.47

of an outlier will vary with working dimension d.

In addition, nonparametric estimation depends on an auxiliary parameter, bandwidth h, and its choice – done here by cross validation – is crucial for the performance of a method. As argued in Ronchetti et al. (1997), an outlier can significantly bias results of the least-squares-based CV. For OPG (3), bandwidth h is chosen generally too small: the CV criterion (4) is minimized when the outlier affects as a small number of observations in its neighborhoud as possible, that is, when the bandwidth h is small. For MAVE (2), the situation is again complicated by the fact that the outlier can be "isolated" not only by a small bandwidth, but possibly by a specific choice of directions B as well. Furthermore, since CV is also used to determine the dimension of the DR space, an outlier can adversely affect the estimation of dimension D as well.

To exemplify the influence of a single outlying observation on the bandwidth and dimension selection, we generated a random sample of 100 observation

Table 2

Optimal CV dimension d for a data set with one additive outlier c.

Simulated data with an additive outlier							
Method	$c = 0$	$c = 25$	$c = 50$	$c = 75$	$c = 100$	$c = 125$	$c = 150$
OPG	1	2	2	3	5	4	1
MAVE	2	2	2	1	3	2	2

from the following nonlinear model:

$$y_i = (X_i^\top b_1)^2 - (0.5 + X_i^\top b_2)^2 + 15\cos(X_i^\top b_3) + 0.5\varepsilon_i,$$

where random vector X_i has the standard normal distribution in $\mathbb{R}^{10}$ (see Section 4 for detailed model description and Monte Carlo simulations). Additionally, we included one observation that has value y_i increased by c ranging from 0 to 1000. We first applied OPG since it allows to determine the local linear approximation (3) and the corresponding bandwidth h separately from directions B. For various values of outlier c and dimension d, the optimal bandwidth h_{opt} was chosen by CV, see Table 1 (line $c = 0$ corresponds to data without the outlier). Although it does not change monotonically with c, there is a general trend of h_{opt} being smaller with increasing outlier value c. Further, we also estimated dimension d as a function of outlier value c. The results of CV based on MAVE and OPG estimates are summarized in Table 2. OPG seems rather sensitive to this outliers because the estimated dimension varies between 1 and 5 for $c = 0, \ldots, 150$. MAVE results in more stable estimates, which are however still influenced by the outlier position.

3 Robust dimension reduction

Both MAVE and OPG seem to be sensitive to data contamination. Our aim is thus to propose robust enhancements of MAVE and OPG that should preserve their present qualities, increase their robustness, and be computationally feasible. We discuss first general ways of making MAVE and OPG more robust (Section 3.1). Next, we address computational issues and propose robust MAVE and OPG that are computationally feasible (Section 3.2). Finally, we adapt the CV procedure mentioned in Section 2.3 for robust estimation (Section 3.3).

3.1 Robust MAVE and OPG

There are many robust alternatives to least squares in linear regression. Using them methods in nonparametric regression adds a requirement on easy and fast implementation, which excludes many so-called high-breakdown point methods (Rousseeuw and Leroy, 2003), and on the other hand, eliminates need for robustness against leverage points to some extent. During last decades, local L- and M-estimators have become particularly popular and well studied, see Boente and Fraiman (1994), Čížek (2004), and Fan and Jiang (1997), Härdle and Tsybakov (1988), respectively.

The application of local L- or M-estimation to MAVE and OPG theoretically reduces to replacing the squared residuals in (2) and (3) by a general convex function ρ_σ of residuals, where σ represents the variance of residuals. Whereas L-estimators do not require the knowledge of σ, in the case of M-estimators, robust choices of ρ_σ depend on an estimate $\hat{\sigma}$ of residual variance. In para-

metric estimation (Hampel et al., 1986), σ is typically treated as a nuisance parameter or is set proportional to the median absolute deviation (MAD). To reflect local character of estimation given by w_{ij} in (2) and (3) and to minimize computational cost, we propose to estimate the variance of residuals by weighted MAD with weights w_{ij}. Specifically, we define

$$\hat{\sigma}(X_0) = 1.4826 \cdot \min_{k=1,\ldots,n} \left\{ r_{(k)} \left| \sum_{i=1}^{n} \frac{K_h(X_i - X_0)}{\sum_{i=1}^{n} K_h(X_i - X_0)} \cdot I\left(r_i \leq r_{(k)}\right) \geq 0.5 \right. \right\},$$

where $r_{(k)}$ is the kth order statistics of $r_i = |y_i - \tilde{\mu}_y(X_0)|$.

In the case of OPG, this means that one applies a local polynomial L- or M-estimator with a given function ρ_σ and a variance estimate $\hat{\sigma}$,

$$\min_{a_j, b_j} \sum_{i=1}^{n} \rho_{\hat{\sigma}}[y_i - \{a_j + b_j^\top (X_i - X_j)\}]^2 w_{ij}, \tag{5}$$

and then obtains the DR space B as d eigenvectors of $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \hat{b}_j^\top \hat{b}_j$ with largest eigenvalues.

In the case of MAVE, having a general objective function $\rho_{\hat{\sigma}}$ leads to

$$\min_{\substack{B: B^\top B = I_p \\ a_j, b_j, j=1,\ldots,n}} \sum_{i=1}^{n} \sum_{j=1}^{n} \rho_{\hat{\sigma}}\left[y_i - \{a_j + b_j^\top B^\top (X_i - X_j)\}\right] w_{ij}. \tag{6}$$

Although (6) cannot be minimized using the algorithm proposed by Xia et al. (2002), the optimization can be still carried out by repeated estimation of (6) with respect to a_j, b_j given an estimate $\hat{B}$ and with respect to B having estimates $\hat{a}_j, \hat{b}_j$. The first step is just the already mentioned local L- or M-estimation analogous to (5). To facilite the second step – estimation of B, let us observe that (6) can be reformulated as follows. For $B = (\beta_1, \ldots, \beta_d)$ and

given $(a_j, b_j) = (a_j, b_{j1}, \ldots, b_{jd})$, (6) is equivalent to

$$\min_{B:B^\top B=I_p} \sum_{i=1}^{n}\sum_{j=1}^{n} \rho_{\hat{\sigma}}\left[y_i - \left\{a_j + \sum_{k=1}^{d} \beta_k^\top b_{jk}(X_i - X_j)\right\}\right] w_{ij}. \tag{7}$$

This represents a standard regression problem with n^2 observations and pd variables, which can be estimated by usual parametric estimator. Although simulations show that estimating MAVE this way leads to slightly better estimates than the original algorithm, the size of regression (7) will be enormous as the sample size increases, which will hinder computation. For example, there are very fast algorithms available for computing least squares and L_1 regression in large data sets, see Koenker and Portnoy (1997), and even in these two special cases computation becomes 10 to 20 times slower than the original algorithm for samples of just 100 observations! This disqualifies such an algorithm from practical use.

3.2 One-step estimation

To be able to employ the fast original MAVE algorithm, robustness has to be achieved only by modifying weights w_{ij} in (2). To achieve this, we propose to use one-step M-estimators as discussed in Fan and Jiang (1999) and Welsh and Ronchetti (2002): the (iteratively) reweighted least squares approach. First using an initial highly robust estimate $\hat{\beta}_0 = \{\hat{B}_0, \hat{a}_{0j}, \hat{b}_{0j}\}$, we construct weights w_{ij}^* such that the objective function (2) is equivalent to (6) at $\hat{\beta}_0$: $w_{ij}^* = w_{ij}\rho_{\hat{\sigma}}(r_{0i})/r_{0i}^2$ where $r_{0i} = y_i - \{\hat{a}_{0j} + \hat{b}_{0j}^\top \hat{B}_0^\top (X_i - X_j)\}$. Next, we perform the original least-squares-based algorithm using the constructed weights w_{ij}^*. Contrary to Fan and Jiang (1999), we use L_1 regression as the initial robust estimator, which guarantees robustness against outliers and fast computation

(one does not have to protect against leverage points since estimation is done in a local window given by the bandwidth and kernel function).

3.3 Robust cross-validation

The robust estimation of the DR matrix B_0 is not sufficient if dimension d is not known. As indicated by Ronchetti et al. (1997) and the example in Section 2.4, using the CV criterion (4) can lead to a bias in dimension estimation (and bandwidth selection) even if a robust estimator of B_0 is used. Now, due to the local nature of nonparametric regression, we have to "protect" CV primarily against outliers in the y-direction. In this context, the L_1 estimator is highly robust and the same should apply to CV based on L_1 rather than L_2 norm. Thus, we propose to use instead of (4) the L_1 cross validation of Wang and Scott (1994),

$$CV(d) = \sum_{j=1}^{n} \left| y_j - \sum_{i=1, i\neq j}^{n} \frac{y_i K_h\{\hat{B}^\top (X_i - X_j)\}}{\sum_{i=1, i\neq j}^{n} K_h\{\hat{B}^\top (X_i - X_j)\}} \right|, \tag{8}$$

to determine both the optimal bandwidth and dimension D. This procedure is further referred to as CVA instead of CV, which is used only for the L_2 cross validation.

4 Simulations

To study finite sample performance of MAVE, OPG and their modifications proposed in Section 3, we perform a set of Monte Carlo simulations under various distributional assumption. In this section, the used data-generating model is introduced first (Section 4.1). Next, we compare the performance

of all methods in estimating the directions of the DR space (Section 4.2). Finally, we examine the performance of the CV and CVA dimension estimation (Section 4.3).

4.1 Simulation models

Throughout this section, we consider the data-generating model

$$y_i = (X_i^\top \beta_1)^2 - (0.5 + X_i^\top \beta_2)^2 + 15\cos(X_i^\top \beta_3) + 0.5\varepsilon_i, \tag{9}$$

where the vector X_i of explanatory variables has the standard normal distribution in $\mathbb{R}^{10}$ and $\beta_1 = (1, 2, 3, 0, 0, 0, 0, 0, 0, 0)/\sqrt{14}$, $\beta_2 = (-2, 1, 0, 1, 0, 0, 0, 0, 0, 0)/\sqrt{6}$, and $\beta_3 = (0, 0, 0, 0, 0, 0, 0, 1, 1, 1)/\sqrt{3}$. The effective DR space given by $B_0 = (\beta_1, \beta_2, \beta_3)$ has thus dimension $D = 3$. To compare the robust properties of all estimators, we use three error distributions.

(1) The standard normal errors, $\varepsilon_i \sim N(0, 1)$, generate Gaussian data without any outlying observations.
(2) The Student distributed errors, $\varepsilon_i \sim t_1$, with one degree of freedom simulate data from a heavy-tailed distribution.
(3) The contaminated errors, $\varepsilon_i \sim 0.95N(0, 1) + 0.05U(-600, 600)$, represent (normal) data containing 5% of outlying observations.

For the sake of brevity, we refer to these three cases as NORMAL, STUDENT, and OUTLIERS, respectively. For all simulations from (9), we use sample size $n = 100$ and 100 repetitions (we observed that the results for larger samples sizes, such as $n = 200$, are qualitatively the same as for $n = 100$). Nonparametric smoothing employs the Gaussian kernel in all cases. Bandwidth

Table 3

Average computational times for each method at sample size $n = 100$ and dimension $d = 3$ relative to standard OPG.

Computation times				
	LS	L_1	HUBER	HAMPEL
OPG	1.0	4.6	10.4	12.3
MAVE	7.5	298.7	13.8	14.5

is cross-validated using CVA for the proposed robust methods and using both CV and CVA for the original methods.

Let us note that we compare the methods using the same distance measure of the estimated space $\hat{B}$ and the true space $B_0 = (\beta_1, \beta_2, \beta_3)$ as Xia et al. (2002): $m(\hat{B}, B_0) = \|(I - B_0 B_0^T)\hat{B}\|$ for $d \leq D = 3$ and $m(\hat{B}, B_0) = \|(I - \hat{B}\hat{B}^T)B_0\|$ for $d \geq D = 3$ and ($D = 3$ is the true dimension of the reduced space used in our simulations, whereas d denotes the dimension used for estimation).

4.2 Estimation of dimension reduction space

MAVE, OPG, and their modifications are now compared by means of simulations. The estimators include MAVE defined in (6) and OPG defined in (5) using the following functions $\rho_{\hat{\sigma}}$:

(1) $\rho_{\hat{\sigma}}(x) = x^2$ (least squares)

(2) $\rho_{\hat{\sigma}}(x) = |x|$ (least absolute deviation),

(3) $\rho_{\hat{\sigma}}(x) = \int \operatorname{sgn}(x) \min(|x|, \hat{\sigma}) dx$ (M-estimate with the Huber function)

(4) $\rho_{\hat{\sigma}}(x) = \int \operatorname{sgn}(x) \max\{0, \min(|x|, \hat{\sigma}) - \max(0, |x| - 2\hat{\sigma})\} dx$ (M-estimate

with the Hampel function)

We refer to these functions as LS, L1, HUBER, and HAMPEL. The first choice corresponds to standard MAVE and OPG. The second one relies on the local L-estimation that, in the case of MAVE, has to be performed by slow algorithm based on alternating formulations (6) and (7), see Section 3.1. The last two choices represent MAVE and OPG relying on the local one-step M-estimation as described in Section 3.2.

Before discussing the estimation results, let us recall our concerns about computational speed of each method that motivated use of sub-optimal, but fast OPG and precludes practical use of MAVE based on the L_1 estimator. Given model (9), Table 3 summarized computational times for all methods relative to OPG–LS. The results are to some extent implementation-specific, which does not allow to directly compare optimized least-square methods and unoptimized robust variants. Nevertheless, it is clear that OPG can be computed much faster than MAVE (if the algorithm is properly optimized) and that the general algorithm from Section 3.1 used for MAVE–L1 is too slow for real applications.

Let us first deal with the results concerning OPG and its modifications presented in Table 4. For data NORMAL, all modifications outperform the original method (OPG–LS–CV). It is interesting to note that, in the case of OPG–LS, the CVA criterion performed better than CV. This might be due to a relatively small number of observations relative to the dimension of the data, but our feeling is that this is typical rather than exceptional for many dimension-reduction applications. Nevertheless, even the comparison of OPG–LS–CVA, OPG–HUBER, and OPG–HAMPEL does not reveal significant differences in

Table 4

Median errors of OPG estimates for dimension $d = 3$.

Simulated data NORMAL, STUDENT, and OUTLIERS

Data	Method	$m_0(\hat{\beta}_1)$	$m_0(\hat{\beta}_2)$	$m_0(\hat{\beta}_3)$	$m_0(\hat{B})$
NORMAL	OPG LS CV	0.005	0.147	0.264	0.352
	OPG LS CVA	0.004	0.132	0.208	0.301
	OPG L1	0.004	0.117	0.215	0.279
	OPG HUBER	0.004	0.127	0.227	0.310
	OPG HAMPEL	0.005	0.125	0.201	0.307
STUDENT	OPG LS CV	0.103	0.521	0.663	0.953
	OPG LS CVA	0.096	0.545	0.599	0.953
	OPG L1	0.021	0.454	0.523	0.883
	OPG HUBER	0.016	0.386	0.530	0.821
	OPG HAMPEL	0.013	0.287	0.467	0.743
OUTLIERS	OPG LS CV	0.641	0.607	0.549	0.969
	OPG LS CVA	0.648	0.616	0.587	0.973
	OPG L1	0.012	0.267	0.473	0.722
	OPG HUBER	0.008	0.192	0.369	0.467
	OPG HAMPEL	0.007	0.167	0.324	0.368

Table 5

Median errors of MAVE estimates for dimension $d = 3$.

Simulated data NORMAL, STUDENT, and OUTLIERS

Data	Method	$m_0(\hat{\beta}_1)$	$m_0(\hat{\beta}_2)$	$m_0(\hat{\beta}_3)$	$m_0(\hat{B})$
NORMAL	MAVE LS CV	0.007	0.092	0.095	0.181
	MAVE LS CVA	0.004	0.089	0.094	0.147
	MAVE L1	0.005	0.060	0.090	0.148
	MAVE HUBER	0.005	0.100	0.094	0.164
	MAVE HAMPEL	0.006	0.116	0.153	0.250
STUDENT	MAVE LS	0.316	0.385	0.572	0.910
	MAVE LS CVA	0.252	0.397	0.510	0.910
	MAVE L1	0.020	0.335	0.388	0.683
	MAVE HUBER	0.035	0.284	0.451	0.685
	MAVE HAMPEL	0.039	0.289	0.428	0.633
OUTLIERS	MAVE LS	0.747	0.732	0.664	0.976
	MAVE LS	0.752	0.682	0.680	0.976
	MAVE L1	0.029	0.165	0.221	0.470
	MAVE HUBER	0.014	0.228	0.202	0.416
	MAVE HAMPEL	0.010	0.151	0.176	0.312

the performance of these methods. For data STUDENT, all robust versions of OPG provide similar results, whereas the estimates by original OPG–LS exhibit rather large errors, especially in the first direction β_1. For data OUTLIERS, there is a large difference between non-robust OPG–LS and the robust modifications of OPG, which documents sensitivity of the original OPG estimator to outliers. Although the performance of all robust estimators is relatively similar, OPG–HAMPEL is best due to its full-rejection feature (observations with too large residuals get zero weight).

The simulation results for MAVE are summarized in Table 5. For data NORMAL, we again observe the positive influence of CVA on the original MAVE. All estimates except for MAVE–HAMPEL perform almost equally well. MAVE–HAMPEL provides worst estimates since it fully rejects some data points, which surprisingly did not matter in the case of OPG. For data STUDENT and OUTLIERS, all robust versions of MAVE by far outperform the original MAVE, which exhibit rather large errors in all direction. Similarly to OPG, MAVE–HAMPEL is best due to the full rejection of extreme observations; this is effect is rather pronounced in data OUTLIERS.

All presented results clearly document the need for and advantages of the proposed robust modifications of MAVE and OPG. Comparing the results for both methods, they have rather similar structure, but MAVE always outperforms OPG when considering the estimation of the whole DR space. On the other hand, OPG seems to be very good in identifying the first and to some extent also the second direction. Let us note that results can be further improved by adjusting function $\rho_{\hat{\sigma}}$, the choice of which was typical rather than optimal.

4.3 Cross-validation simulations

The estimation of the DR space considered in the previous section is now complemented by a study on the estimation of the effective dimension D. Simulating again from model (9), we estimated the DR dimension $\hat{d}$ ($D = 3$) using MAVE and OPG with $\rho_{\hat{\sigma}}$-functions LS and HAMPEL and the CV and CVA criteria (4) and (8), respectively. Note that the design of model (9) makes the identification of the third direction β_3 difficult given rather small sample size $n = 100$. Therefore, we accept estimates $\hat{d} = 2$ and $\hat{d} = 3$ as appropriate.

Results for all models are summarized in Tables 6 and 7 for MAVE and OPG, respectively. Judging all methods by the number of simulated data sets for which estimated dimension equals two or three, MAVE can be always preferred to OPG. For the original methods, the CVA criterion is preferable to CV as in Section 4.2. A more interesting results is that MAVE–HAMPEL and OPG–HAMPEL outperformed the original MAVE and OPG not only in data OUTLIERS and STUDENT, but also in the case of data NORMAL. The only case where MAVE–LS is preferable is the number of data set for which the estimated dimension $\hat{d}$ equals 3 in data NORMAL. This could be partially accounted to the relatively small sample size. Finally, notice that the robust DR method such as MAVE–HAMPEL does not suffice to identify the dimension of the DR space: a robust CV criterion such as CVA has to be used as well.

5 Conclusion

We proposed robust enhacements of MAVE and OPG that can perform equally well as the original methods under 'normal' data, are robust to outliers and

Table 6

Estimates of the DR dimension by MAVE variants using L_2 (CV) and L_1 (CVA) cross validation. Entries represent the numbers of samples out of 100 with estimated dimension d.

Simulated data NORMAL, STUDENT, and OUTLIER

Data	Estimation	CV	$D = 1$	$D = 2$	$D = 3$	$D = 4$	$D \geq 5$
NORMAL	MAVE LS	CV	11	38	51	0	0
NORMAL	MAVE LS	CVA	4	43	53	0	0
NORMAL	MAVE HAMPEL	CV	7	52	41	0	0
NORMAL	MAVE HAMPEL	CVA	6	54	40	0	0
STUDENT	MAVE LS	CV	26	33	35	5	1
STUDENT	MAVE LS	CVA	15	29	44	12	1
STUDENT	MAVE HAMPEL	CV	18	42	27	7	5
STUDENT	MAVE HAMPEL	CVA	15	62	21	2	0
OUTLIERS	MAVE LS	CV	0	1	14	20	65
OUTLIERS	MAVE LS	CVA	5	6	20	10	59
OUTLIERS	MAVE HAMPEL	CV	1	2	6	10	81
OUTLIERS	MAVE HAMPEL	CVA	8	47	29	4	12

Table 7

Estimates of the DR dimension by OPG variants using L_2 (CV) and L_1 (CVA) cross validation. Entries represent the numbers of samples out of 100 with estimated dimension d.

Simulated data NORMAL, STUDENT, and OUTLIER

Data	Estimation	CV	$D = 1$	$D = 2$	$D = 3$	$D = 4$	$D \geq 5$
NORMAL	OPG LS	CV	8	48	38	5	1
NORMAL	OPG LS	CVA	8	59	31	1	1
NORMAL	OPG HAMPEL	CV	8	51	40	1	0
NORMAL	OPG HAMPEL	CVA	8	47	43	2	0
STUDENT	OPG LS	CV	21	29	27	18	5
STUDENT	OPG LS	CVA	11	25	41	18	5
STUDENT	OPG HAMPEL	CV	10	31	33	14	12
STUDENT	OPG HAMPEL	CVA	23	41	31	5	0
OUTLIERS	OPG LS	CV	2	1	10	12	75
OUTLIERS	OPG LS	CVA	14	5	8	8	65
OUTLIERS	OPG HAMPEL	CV	0	0	2	13	85
OUTLIERS	OPG HAMPEL	CVA	12	40	23	10	15

heavy-tailed distributions, and are easy to implement. Should we pick up one method for a general use, MAVE–HUBER seems to be the most suitable candidate as (i) MAVE–LS is not robust, (ii) MAVE–L1 is slow to compute, see Section 3.1, and (iii) MAVE–HAMPEL does not perform so well for normal data.

References

[1] Amato, U., Antoniadis, A., and De Feis, I., 2005. Dimension reduction in function regression with applications. Computational Statistics & Data Analysis, in press.

[2] Antoniadis, A., Lambert-Lacroix, S., and Leblanc, F., 2003. Effective dimension reduction methods for tumor classification using gene expression data. Bioinformatics 19(5), 563–570.

[3] Boente, G. and Fraiman, R., 1994. Local L-estimators for nonparametric regression under dependence. Journal of Nonparametric Statistics 4, 91–101.

[4] Čížek, P., 2004. Smoothed local L-estimation with an application. In: M. Hubert, G. Pison, A. Struyf, and S. Van Aelst (Eds.), Theory and Applications of Recent Robust Methods, Birkhäuser, Basel, 59–70.

[5] Fan, J. and Jiang, J., 1999. Variable bandwidth and one-step local M-estimator. Science of China, Ser. A, 29, 1–15.

[6] Gather, U., Hilker, T., and Becker, C., 2001. A robustified version of sliced inverse regression. In: L. T. Fernholz, S. Morgenthaler, and W. Stahel (Eds.), Statistics in Genetics and in the Environmental Sciences, Birkhäuser, Basel, 147–157.

[7] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A., 1986. Robust statistics, the approach based on influence function. Wiley, New York.

[8] Härdle, W. and Stoker, T. M., 1989. Investigating smooth multiple regression by method of average derivatives. Journal of American Statistical Association 84, 986–995.

[9] Härdle, W. and Tsybakov, A. B., 1988. Robust nonparametric regression with simultaneous scale curve estimation. Annals of Statistics 16, 120–135.

[10] Koenker, R. and Portnoy, S., 1997. The Gaussian hare and the Laplacian tortoise: computability of squared-error vs. absolute-error estimators. Statistical Science 12, 279–300.

[11] Li, K. C., 1991. Sliced inverse regression for dimension reduction. Journal of American Statistical Association 86, 316–342.

[12] Rousseeuw, P. J. and Leroy, A. M., 2003. Robust regression and outlier detection. Wiley, New York.

[13] Ronchetti, E., Field, C., and Blanchard, W., 1997. Robust linear model selection by cross-validation. Journal of American Statistical Association 92, 1017–1023.

[14] Sakata, S. and White, H., 1998. High breakdown point conditional dispersion estimation with application to S&P 500 daily returns volatility. Econometrica 66(3), 529–567.

[15] Wang, F. T. and Scott, D. W., 1994. The L_1 method for robust nonparametric regression. Journal of American Statistical Association 89, 65–76.

[16] Welsh, A. H. and Ronchetti, E., 2002. A journey in single steps: robust one-step M-estimation in linear regression. Journal of Statistical Planning and Inference 103, 287–310.

[17] Xia, Y., Tong, H., Li, W. K., and Zhu, L.-X., 2002. An adaptive estimation of dimension reduction space. Journal of Royal Statistical Society, Ser. B, 64, 363–410.

[18] Yao, Q. and Tong, H., 1994. On subset selection in nonparametric stochastic regression. Statistica Sinica 4, 51–70.